A CARTOGRAPHIC OPTICAL CHARACTER RECOGNITION SYSTEM

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfilment of the Requirements

For the Degree of

Master of Science

in the

Department of Electrical Engineering

University of Saskatchewan

by

Ron Bolton

Saskatoon, Saskatchewan

May 1977

ACKNOWLEDGEMENTS

The author would like to express his sincere thanks to Dr. A.R. Boyle whose ideas, advice and guidance during the project proved invaluable.

Thanks are also due to the Graphic Systems Design and Applications Group at the University of Saskatchewan. In particular, the author would like to thank Mr. A. Schiller for help with the video quantizer board and Mr. K. Bourassa whose ability to transform ideas into computer routines was excellent.

Finally, the author would like to express his gratitude to the United States Naval Oceanographic Office for making it possible to undertake this work.

This work was supported by NAVOCEANO contract #N62306-74-C-0006.

UNIVERSITY OF SASKATCHEWAN
ELECTRICAL ENGINEERING ABSTRACT 77A178

'A CARTOGRAPHIC OPTICAL CHARACTER RECOGNITION SYSTEM'

Student: Ron Bolton Supervisor: A.R. Boyle

M.Sc. Thesis presented to College of Graduate Studies

May 1977

ABSTRACT

This thesis describes an Optical Character Recognition system for use in a cartographic application. The system is primarily intended to recognize and digitize both machine printed navigational chart sounding values and hand-printed field sheet sounding values.

The system consists of a precision X-Y flatbed transport, a vidicon camera, a PDP 8/e minicomputer, an interactive display console and associated software and hardware.

After investigating numerous recognition algorithms, three were finally chosen to be included in the system. The final recognition decision was decided by a majority vote. An interactive sounding editing capability was included so that 'unrecognized' soundings could be input manually by the program operator.

The complexity of the problem necessitated the construction of sophisticated hardware for both the X-Y transport and the vidicon camera. As well, complete input and output software routines were included to make the system fully integrated.

Experimental results are given for the system.

This research was supported by the United States Naval Oceanographic Office.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
LIST OF FIGURES	viii
LIST OF TABLES	xi
LIST OF PHOTOGRAPHS	xii
LIST OF FLOWCHARTS	xiii
1. INTRODUCTION	1
1.1 Optical Character Recognition (OCR) in General	1
1.2 University of Saskatchewan OCR Systems	3
2. A BASIC OCR SYSTEM	9
2.1 General	9
2.2 Input	11
2.3 Transport	11
2.4 Scanner	12
2.4.1 Flying Spot Scanner	12
2.4.2 Vidicon Raster Scanner	12
2.4.3 Line or Array Scanner	12
2.5 Preprocessing	13
2.5.1 Enhancement	13
2.5.2 Transformation	13

Table of Contents (Cont'd)

	Page
2.6 Recognition	14
2.6.1 Entire Character (Gestalt)	15
2.6.2 Essential Feature Extraction	15
2.6.3 Threshold Logic (Decision Functions)	16
2.6.4 Information Transformation	16
2.7 Output	17
3. THE NAVOCEANO/UNIVERSITY OF SASKATCHEWAN OCR SYSTEM	18
3.1 General	18
3.2 Acquisition of Images	18
3.3 Extraction and Enhancement	20
3.4 Recognition of Characters	20
3.5 Output of Data	21
3.6 Explanation of Development	21
4. SPECIAL FEATURES OF THE NAVOCEANO/UNIVERSITY OF SASKATCHEWAN OCR SYSTEM	23
4.1 General	23
4.2 Program Swapping	23
4.3 Image Storage and Recall	25
4.4 Chart Storage and Recall	28
4.5 Standards File Specification	29
4.6 Problem Areas Specification	29
5. ACQUISITION OF IMAGES	30
5.1 The Input Material	30
5.2 The Transport System	32

Table of Contents (Cont'd)

	Page
5.2.1 Rotating Drum	32
5.2.2 Flatbed X-Y Plotter	34
5.3 The Scanner System	37
5.3.1 The Quantizer and Synchronizing Board	41
6. EXTRACTION AND ENHANCEMENT	46
6.1 General	46
6.2 Preprocessing (Software)	46
6.3 Normalization	50
6.3.1 Padded Array Normalization	51
6.3.2 Aspect Ratio Normalization	53
7. RECOGNITION	56
7.1 General	56
7.2 The Grid Method	58
7.3 The Template Method	63
7.4 The Characteristic Waveform Method	68
8. OUTPUT OF DATA	75
8.1 Postprocessing	75
8.2 Data Output	76
8.2.1 Mainheader	81
8.2.2 Subheaders	81
8.2.3 Control Points	82
8.2.4 Check Points	82
8.2.4.1 Main Chart Boundary	82
8.2.4.2 Problem Area Boundary	83

Table of Contents (Cont'd)

		Page
8.2.5	Soundings	83
9.	OPERATOR-COMPUTER INTERACTION	86
9.1	General	86
9.2	Special Chart Information Entry	86
9.3	Standards File Specification	88
9.4	Problem Areas Information Entry	93
9.5	Computer Editing Facilities (soundings)	94
9.6	Computer Console Switch Register Options	95
10.	EXPERIMENTAL RESULTS	97
11.	RECOMMENDATIONS	106
11.1	The Histogram Estimated Probability Density Method	108
12.	CONCLUSIONS	115
13.	REFERENCES	117

APPENDICES

APPENDIX A	Calculation of the Weighted Centre of a Sounding	119
APPENDIX B	OCR User's Manual	123
APPENDIX C	Data File Formats	140
	C.1 Data File Specification	141
	C.2 Data Formats (Disk)	141
APPENDIX D	Quantizer and Synchronizing Board Schematics	157
APPENDIX E	Examples of Different Fonts	174
APPENDIX F	Individual Character Representations for Each of the Three Recognition Methods	179

LIST OF FIGURES

Figure		Page
2.1	Basic OCR System Block Diagram	10
3.1	NAVOCEANO/UNIVERSITY OF SASKATCHEWAN OCR System Block Diagram	19
4.1	OCR Program Module Functions	24
4.2	OCR Program Core Map	26
4.3	OCR Program Swapping File Structure	27
5.1	Table Movement (Major)	38
6.1	Examples of Illegal Soundings	49
6.2	Padded Array Normalization	52
6.3	Aspect Ratio Normalization	54
6.4	Aspect Ratio Normalization -- Numerals 0-9	55
7.1	Grid Method Scan Lines	60
7.2	Grid Method Transition Table	60
7.3	Template Recognition Method	65
7.4	Characteristic Waveform Method	69
8.1	Weighting Example #1	77
8.2	Weighting Example #2	78
8.3	Weighting Example #3	79
8.4	Weighting Example #4	80

List of Figures (Cont'd)

Figure		Page
9.1	OCR Program Operator Information (Standards File)	90
11.1	Probability (Master) Matrix	111
11.2	Histogram Estimated Probability Density Results	113
11.3	Training Data for One Class	114
A.1	Calculation Example	121
A.2	Weighted Center Examples	121
B.1	Bit Location on Console	134
C.1	Overall Data Format	142
C.2	General Data Structure	144
C.3	Data Directory Format (One Entry)	146
C.4	Subheader Format	147
C.5	Binary Multipliers	148
C.6	Data Code 2 Format	150
C.7	Tags	151
C.8	Continuous Format	151
C.9	Data Code 6 Format	152
C.10	Rectangular Area (Check Points)	153
C.11a	Subheader Block Format (Data Code 10)	154
C.11b	Data Block Format (Data Code 10)	154
C.12	OS/8 Data Word	155
C.13	Word Grid	156

List of Figures (Cont'd)

Figure		Page
D.1	Instructions	158
D.2	Power Clear Buffer	159
D.3	+12V,,-6V Power	160
D.4	High Frequency Logic	161
D.5	Low Frequency Logic	162
D.6	CCU Logic Changes	163
D.7	Sync. Level Shifters	164
D.8	Video Switcher	165
D.9	Video Amplifier	166
D.10	Video Shifter	167
D.11	V_{agc} Logic	168
D.12	Video Clamp	169
D.13	Video Quantizer	170
D.14	Box Shader	171
D.15	Video Mixer	172
D.16	IC Locations	173
E.1	USN Map Font	175
E.2	OCR B Font	176
E.3	Hand Printed Font #1	177
E.4	Hand Printed Font #2	178
F.1	Numerals 0-9 (Normalized)	180
F.2	Grid Method Representation	181
F.3	Template Method Representation	182
F.4	Characteristic Waveform Method Representation	183

LIST OF TABLES

Table		Page
10.1	Standard Test Recognition Results	98
10.2	Individual Recognition Method Results	99
10.3	Digit and Individual Recognition Method Rates	100
10.4	Final Recognition Rates	101
10.5	'Standards File' Recognition Results	102
10.6	'Ideal File' Recognition Results	102

LIST OF PHOTOGRAPHS

Photograph		Page
5.1	Example of a Hand Printed Chart	31
5.2	Example of a Machine Printed Chart	31
5.3	Gerber 22 Controller	36
5.4	Gerber 22 X-Y Transport	36
5.5	Camera Head Shown Mounted on Gerber 22 Gantry ...	40
5.6	Quantizer and Synchronizing Board	42
5.7	Databreak Control Board	42
5.8	Sampling Board	43
5.9	Master Control Board	43
5.10	Example of Quantized Video Output	44
6.1	Example of an Image with Data Clutter	48
9.1	PDP 8/e Minicomputer and 4015 Display Console ...	87
9.2	Display Format	96

LIST OF FLOWCHARTS

Flowchart		Page
7.1	The Grid Method	61
7.2	The Template Method	66
7.3	The Characteristic Waveform Method	70

1. INTRODUCTION

1.1 Optical Character Recognition (OCR) In General

Pattern recognition, an essentially new field of technology, has been extensively investigated during the last twenty years. It may be defined as the extraction of significant information from a background of noise, and is widely used in such diverse fields as speech recognition, medical analysis (electroencephalograms and electrocardiograms) and more recently, in image processing.

Image processing, as a subset of pattern recognition, involves the study of both two and three dimensional images. Three dimensional images, ones which retain grey levels of the original, include reconnaissance and weather photographs, medical X-rays and most recently, Earth Resources Technology Satellite (ERTS) and Landsat image data. Two dimensional images, ones in which the grey levels are quantized (or binarized) to black and white, include fingerprints, map reading and character recognition. The topic of this thesis will be two dimensional optical character recognition, or OCR as it is known.

Character recognition is the transcription of humanly legible alphabetic and numeric information to some other form required for a communication-transmission system. This 'other form' may be mechanical (as in the OPTICON$_{TM}$ reader for the blind), but more usually is a digital form which can be stored, encoded, translated

or manipulated using a digital computer.

This transcription is not an easy task. Generally commercial models of OCR equipment have taken a straightforward approach to the problem and the results have varied from poor to good depending on the application. Systems are very expensive (almost to the point of being prohibitive) and the information input has usually been constrained (i.e. within a bounding box). More recent versions of OCR systems have tried more elegant means of accomplishing the task. They are very flexible, allowing the processing of multiple fonts; surprisingly the cost has usually been lower than in the earlier versions. There has been, however, a gap between the ideas of manufacturers of the OCR equipment and the thinking of the researchers. This gap is now shrinking rapidly and more and more of the commercial systems are incorporating some of the immense amount of research that has been conducted in the OCR field.

The OCR problem was at first approached with the optimism that is characteristic of this technological age. This optimism led to predictions which were impossible to fulfill. As a result, OCR has developed the reputation of being a futuristic consideration. Only recent developments, coupled with a cautious optimism and the realization that any one OCR is not going to be a final answer to the problem, has again placed OCR systems with those which may be considered viable.

The basis for the large amounts of research and funding being directed to the character recognition problem, must now be considered.

The main force behind this activity may be termed the 'information explosion'. When the number of bank cheques, government forms and pieces of mail to be checked, indexed and sorted, approaches several billions each week, man then requires the help of machines. This large amount of information is also accompanied by the need to process such information faster than ever before. In other words, the increase in information is accompanied by the need for a decrease in the access time to this information. There are, in fact, a large number of companies involved in not only the production of OCR equipment, but also in the use of OCR equipment for processing information.

As an illustration of the above problem, it is estimated that man's total accumulation of knowledge doubled during the seven years from 1960 to 1967[1][2]. Previously, it had doubled during the 250 years from 1750 to 1900, doubled again the 50 years from 1900 to 1950, and again in the 10 years from 1950 to 1960.

This thesis reports on the progress made at the University of Saskatchewan with OCR as an aid to computerized cartography.

1.2 University of Saskatchewan OCR Systems

In recent years the need for an efficient means of reproducing maps and nautical charts has been investigated at the University of Saskatchewan by the Graphic Systems Design and Applications Group under the direction of A.R. Boyle. As mentioned previously, the need to access larger and larger amounts of data at higher and higher speeds has necessitated a digital data bank approach. This

is especially important in the cartographic area since, in order to reporduce a map of chart accurately, it is necessary to manipulate large amounts of data to a high degree of precision.

A further impetus to automate is the recent decision of both Canada and the United States to convert to metric measurements, which means that most maps and nautical charts must now be redrawn.

One product of years of research at the University of Saskatchewan has been the interactive cartographic system, CAMC (Computer Aided Map Compilation), which allows cartographers to manipulate map and chart data in an efficient manner.

Some interesting auxiliary aspects of the CAMC system are:

1) automatic drafting on film[3]

2) interactive digitization of maps and charts[4]

In the interests of trying to expand the CAMC system, work was started in 1969 on an automatic optical character recognition (OCR) system for use with nautical charts. The nautical charts (maps used for navigation or to delineate marine areas) contain many soundings (numerical values which indicate the depth to bottom at their location on the chart) which could be recognized automatically. The original work was done by S.K. Agarwal supported by funds from the Canadian Hydrographic Service[5][6]. His work proved that the fundamental approach to the recognition of nautical soundings, using a high precision plotting table as a transport mechanism and a vidicon camera as a scanner, was feasible. However, the system had the following defects:

1) the picture data transfer rates were very slow

2) the chart data sheets had to be pre-processed by hand to remove lines and 'data clutter' before recognition could be attempted.

3) an inability to process 'suffix' digits and special symbols

4) an inability to process hand-printed characters

5) no provision for tagging or editing unrecognized soundings was provided

6) the system was not integrated, as no provision was made for special information entry, data output (soundings and coordinates) or adaptive-interactive operations.

While the defects of the system were obviously beyond the scope of the original development, further work to produce an integrated system was needed. A second generation system was, therefore, developed by R. Brooks in 1970. It attempted to recognize hand-printed soundings and to output the resulting data (soundings and coordinates). However, work was halted on this system in 1972.

The Graphics System Design and Applications Group was then approached in 1973 by the United States Naval Oceanographic Office (USNOO-NAVOCEANO) with a specification[7] for a total nautical chart sounding recognition system. Their need for such a system was more pressing than that of the Canadian Hydrographic Service as they were in the process of attempting to automate the entire system of sounding data acquisition. In the interim they needed a system or systems, which were able to digitize the various nautical charts and field sheets in use at the time.

The specification for the NAVOCEANO sounding recognition system included:

1) rejection rate less than 1 percent for printed (optical or mechanical) numerics; less than 25 percent for hand-printed numerics
2) substitution rate less than 0.1 percent
3) digitization and recognition of one complete sounding in one second
4) resolution, ±0.002"; repeatibility; ±0.002", accuracy, ±0.007"
5) chart sizes up to 48 inches by 58 inches must be handled
6) all 'unrecognized' digits or symbols must be tagged for later manual correction
7) complete sounding value and position editing procedures must be included
8) must be compatible with CAMC system data formats
9) ability to store and retrieve pictures to/from an industry compatible magnetic tape
10) must include a diagnostic routine to check picture linearity, size and table repeatibility.
11) ability to process suffix digits
12) ability to process special characters (6)
13) limited ability to remove 'data clutter' such as lines, smudges and large symbols from picture data.

The purpose, then, of this research was to develop a minicomputer based OCR system for use in an interim cartographic application, until the new, fully automated sounding data acquisition system can into production. The unique nature of the application presented some difficulties. For instance it would be used to

process both existing printed nautical charts, and also the USN field sheets which are hand-printed. This necessitated a multiple character set capability which could only be accomplished by using interactive processes[23]. It is estimated that the USN field sheets number in the tens of thousands and thus it is imperative that the system be able to process the approximately four foot by five foot sheets as rapidly as possible. For this reason, special computer software and hardware were incorporated into the OCR system. The specifications for both position accuracy and substitution rate (a wrong decision for a sounding digit) are both extremely important, as the country responsible for the final charts is liable in event of a maritime disaster. The difficulty of obtaining a low substitution rate was eased due to the fact that only sixteen different characters (the ten digits and six special symbols) would be encountered, rather than a full alphanumeric character set. Three relatively simple and easily programmed recognition algorithms with a weighted majority vote were thus able to produce the final recognition decision. This weighting scheme and the use of interactive processes at some stages of the chart processing, enhanced the reliability of the system recognition rates significantly. The OCR program was developed to be a modular and flexible one, so that the present recognition methods may be replaced with newer and possibly more applicable ones, whenever needed. This flexibility also enables the potential users of the system to tailor it to the particular type of chart being processed (i.e. charts where tilted digits are encountered).

This thesis will show that the above purposes were in fact satisfied by the present system. The author is at present working on a more efficient recognition algorithm that, if successful, will later be incorporated into a new version of the system.

The rest of this thesis is devoted to a detailed discussion of the research and development carried out by the author. First a basic OCR system consisting of six main elements (input, transport, scanner, preprocessor, recognition and output) is discussed. The proposed cartographic OCR system is then discussed in general, and some special features are also mentioned which, while not directly related to OCR, have proved invaluable in the development phase of the research. The resulting OCR systems is then described using four main areas of discussion (image acquisition, extraction and enhancement, recognition or characteristics and output of data). The system design is followed by a chapter on the interactive processes necessary to insure the integrity of the data and also to reduce the sophistication needed in the computer program. Experimental results are followed by recommendations to improve the system performance. Included in these is a more efficient method of recognition for use with hand-printed characters. Appendices are also included to demonstrate the amount of integration necessary to interface this OCR system efficiently into an existing cartographic computer system.

2. A BASIC OCR SYSTEM

2.1 General

The basic character recognition problem may be stated as follows[8]:

> Optical character recognition requires unique and unequivocal identification of two-dimensional signal structures which comprise ideal alphanumeric characters embedded in noise. The noise generally consists of stylistic variations and random combinations of dilation, translation and rotation contaminated by superimposed and deleted regions of varying spatial extent. All of this is subject to a continuous range of signal strength and contrast.

Thus, a problem which at first glance appears to be straightforward and simple, mainly because of our preconditioning due to years of experience, can be in reality intricate and expensive.

The types of OCR systems in use at the present time range from simple units with limited capability to complex, expensive systems which are fast and accurate recognition units.

The block diagram of a basic optical character recognition system is shown in Fig. 2.1.

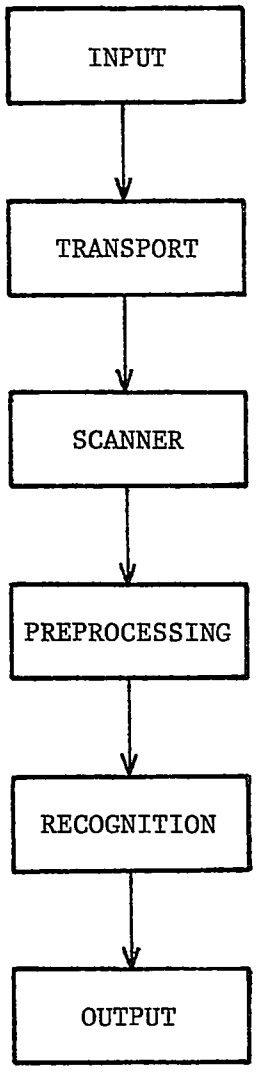

Fig. 2.1 Basic OCR System Block Diagram

2.2 Input

The input into any OCR system is the material base carrying the characters to be recognized. This material base may be a paper roll (cash register tally), a document (bank cheque), a page (book), or even a map (hydrographic chart).

The general type of input usually classifies the recognition system as a roll reader, a document reader, a page reader, or a chart reader.

The form of the information encountered on the input material significantly influences the amount of sophistication required in the OCR system. The information may contain mark sense, machine print or hand-printing. The size of the alphabet may vary from numeric only to alphabetic-numeric plus special characters and may include multiple-font characters (for instance, characters from different typewriters).

2.3 Transport

The transport system physically moves the input material through the recognition system. In a roll reader the roll is moved past the scan station, then wound on a take-up roll. Document and page readers are more complex because they must select a single document from the input stack, align it, feed it past the scan station (where it may be either stopped for scanning or scanned on the fly), and then restack the document. In a chart reader the major concern is knowing the exact location of the character on the chart. This necessitates a high precision X-Y transport system and as a result, the rate of travel of the chart past the scan station (in certain systems the scan station is moved and the chart is held stationary) is very low.

2.4 Scanner

The optical scanner is a device which converts the light reflected from, or transmitted through, the input material into a signal or a series of signals which can be used by the preprocessing and recognition units.

A few of the more common types of scanners are discussed below.

2.4.1 Flying Spot Scanner

With this scanner a controlled, localized beam of light can scan up and down and across the character in a raster fashion. The amount of reflected light is measured by a light sensitive device as a function of time.

2.4.2 Vidicon Raster Scanner

With this scanner the document (or chart) field is uniformly illuminated and the reflected light is temporarily retained by a vidicon target to be read out at a later time in a raster format (television). One of the main advantages of this method is that the vidicon target (storage medium) acts as an integrator and, thus, the signal to noise ratio is usually better than in a flying spot scanner.

2.4.3 Line or Array Scanner

With this type of scanner system a matrix of semiconductor elements (either 1 x n or m x n) measures the amount of light reflected off the surface of the input material and stores the information until it can be read. It is, in effect, a solid state vidicon target. This type of scanner is relatively

new and until recently the size of such scanners has been limited to approximately 50 elements by 50 elements which could often severely restrict the speed performance of the system.

2.5 Preprocessing

The preprocessor 'processes' scanner output in a manner necessary to ease the load of the recognition unit. Basically it tries to remove all variability within the same pattern class and at the same time, increase the variability between different pattern classes.

There are two different types of preprocessor functions. They are:

1) enhancement
2) transformation

2.5.1 Enhancement

Enhancement functions take raw scanner data input and keeping the basic scanner input form of the data, process it so that the output is in an enhanced form. Examples of enhancement functions are size normalization, stroke width standardization, registration, de-skewing, and filtering. Another important function is picture segmentation or character separation, at which time 'data clutter' such as lines may be removed.

2.5.2 Transformation

Transformation functions take raw scanner data input

and changing the basic scanner input form of the data to another form of input, process it so that the output is in transformed form. Examples of transformation functions are feature lists, moments, correlation functions, and feature measurements.

The type of input material dictates not only the pre-processing functions which are needed, but also the complexity and effectiveness of each category -- enhancement or transformation. For example, in a multiple-font processing system (one which involves style, shape and size variations) enhancement techniques are not satisfactory and transformation functions are more appropriate.

2.6 Recognition

The recognition techniques used in the recognition unit may be classified in many ways. Some are:

1) adaptive
2) fixed
3) supervised adaptive
4) non-supervised adaptive
5) statistical
6) software oriented
7) serial
8) parallel
9) etc.

The recognition technique(s) selected are influenced very substantially by the type of input to be presented: single-font,

multiple-font, hand-printed, numeric only, alphabetic and numeric, alphanumeric and special characters, input material format, print quality, etc., and a large amount of experimentation must first be done before accepting a recognition technique for any particular application.

In general, any recognition technique fits into one of the following four general categories:[1][18]

1) entire character (gestalt) recognition
2) essential feature extraction
3) threshold logic or decision functions
4) information transformation

2.6.1 Entire Character (gestalt) Recognition

Most of the gestalt methods of character recognition are optical. The character to be read is illuminated and the amount of light reflected from it through a set of templates is measured. This method can become very complicated if processing a full alphanumeric character set because of the similarity between certain classes (F and E, O and Q, etc.). The method is fairly easy to fabricate; however, it requires at least one template for each character.[8]

2.6.2 Essential Feature Extraction

Any object or pattern (character) which can be recognized and classified possesses a number of discriminatory properties or patterns. The main problem is to choose what features to select and how to extract them.[17]

Consider the recognition of hand-printed characters. The more important features are the sequences of strokes, the direction of the strokes, the arrangements of the strokes and the interrelation between the strokes. None of these features are easily measured; however, there are powerful recognition techniques which use these features. A prime example of this is the stroke analysis method[9][10][11][16]. The grid method used by the author is an essential feature extraction method.

2.6.3 Threshold Logic (decision functions)

These types of recognition methods are characterized by a two-stage process[20]. The first stage is a training phase where the samples of each class have measurements taken on them and decision boundaries are made up. These boundaries separate each class from all others. The second phase consists of taking the samples to be recognized and using the decision boundaries as class separation functions group each of the characters into one of the classes.

Examples of this method are the perceptron, successive dichotomy and, in certain instances, the template method.

2.6.4 Information Transformation

These methods are characterized by the transformation of the original character measurements to another form on which the recognition process operates. Some information transformation methods are vector crossing, curve following and coefficient analysis[8]. Another method, one which was

used in this OCR system, is the Characteristic Waveform method.

2.7 Output

The output of the recognition system is a string of decisions from the recognition unit. The performance of the entire recognition system is judged by:

1) the correctness of the string of decisions
2) the speed at which the decisions are produced
3) the cost of producing the decisions

Two types of errors are generally acknowledged:

1) substitution errors
2) rejection errors

Of these two, substitution errors are in general the more serious. In a substitution error the wrong decision is given: for example, a 5 is identified as a 6. In a reject error no decision is reached, but an 'unknown' character is acknowledged. In most systems some trade-off is made for decreasing the substitution errors with a corresponding increase in the reject errors. The latter may then be corrected at a later time (by searching the output data for 'unknown character' tags) or even in real time (by use of a CRT and interactively inserting the correct answer). Naturally the usefulness of a trade-off depends significantly on the criticallity of either error. In chart reading, for example, the substitution rate must be kept as low as possible while reasonably high reject rates can be allowed: as a result an interactive editing capability is usually included in a chart OCR system.

3. THE NAVOCEANO/UNIVERSITY OF SASKATCHEWAN OCR SYSTEM

3.1 General

As stated previously, OCR systems may vary from the simple to the complex. The NAVOCEANO/UNIVERSITY OF SASKATCHEWAN OCR SYSTEM (referred to in the remainder of this thesis as the OCR system) had the problem of reading characters from map (or chart) sheets. The problem was complicated by the fact that some of these charts were hand-printed, but simplified by the fact that:

1) the data contained at most 16 different classes of characters (numerals 0 through 9 and six special characters), and,
2) the system would be allowed to output 'unknown' whenever it was unable to reach a confident decision as to the identity of the character being processed (substitution errors must be avoided as much as possible).

The block diagram of the system is shown in Fig. 3.1

3.2 Acquisition of Images

The system had to read characters which had been printed (machine-printed or hand-printed) on a large area input sheet. Therefore, a mechanical transport system and an electronic scanner system, both controlled by a DEC PDP 8/e minicomputer, had first to be developed. The transport system was one with which the University of Saskatchewan had had previous experience (Gerber 22), but the

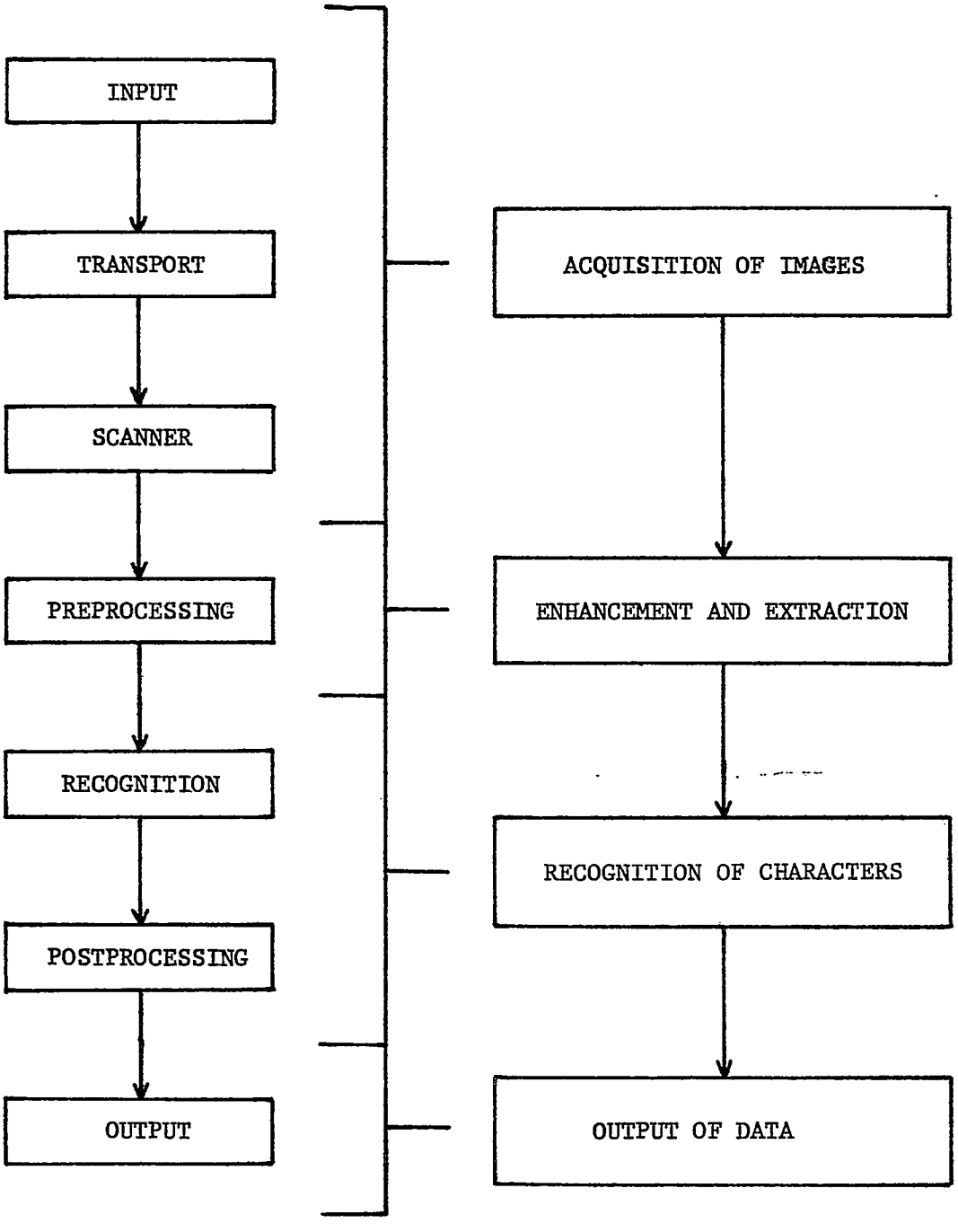

Fig. 3.1 Block Diagram of NAVOCEANO/UNIVERSITY SASKATCHEWAN OCR System

scanner system (Sierra Scientific LV-1) required the design of special electronic controllers in order to function as required.

3.3 Enhancement and Extraction

Enhancement of the images consisted of quantization of the multilevel video picture. This was accomplished using special hardware developed by the author at the University of Saskatchewan.

The extraction section of the OCR program posed a problem because it was necessary to remove extraneous lines, smudges, large symbols, partial soundings, etc., from the enhanced images while leaving the valid characters to be recognized intact. This problem was essentially overcome so that at present only very cluttered areas of the map must be left for manual operator identification. A normalization technique was also used to standardize the size of the characters being input into the recognition unit.

3.4 Recognition of Characters

An examination of the recognition problem showed that the requirements the recognition unit had to meet were very diverse: type of input material base (paper, mylar, etc.), format of the characters on the input material, size of the input material, etc. A literature search indicated that there was no consensus of opinion as to the best method of recognition for this particular application. The system was, therefore, developed to be modular and flexible, allowing the potential user to change parameters of recognition in the future as might seem advisable and be able to develop and add new methods as they became available.